HZ BOOKS
华章教育

Pearson

经 典 原 版 书 库

大数据导论

[美] 托马斯·埃尔 Thomas Erl　瓦吉德·哈塔克 Wajid Khattak　保罗·布勒 Paul Buhler 著

（英文版）



Big Data Fundamentals

Concepts, Drivers & Techniques

Co-authored and Edited by Best-Selling Author Thomas Erl
Co-authored by Wajid Khattak and Paul Buhler, PhD

机械工业出版社
China Machine Press

PRENTICE HALL

ServiceTech PRESS

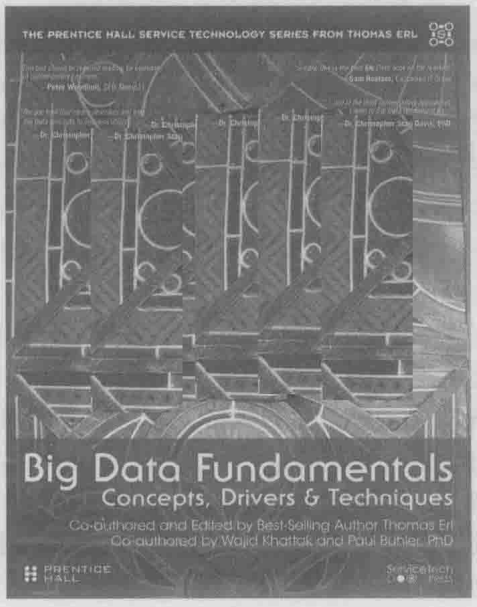
经 典 原 版 书 库

大数据导论

（英文版）

Big Data Fundamentals

Concepts, Drivers, and Techniques



[美] 托马斯·埃尔 瓦吉德·哈塔克 保罗·布勒 著

Thomas Erl Wajid Khattak Paul Buhler

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据导论（英文版）/（美）托马斯·埃尔（Thomas Erl）等著．—北京：机械工业出版社，2017.9
（经典原版书库）
书名原文：Big Data Fundamentals: Concepts, Drivers, and Techniques

ISBN 978-7-111-58098-0

Ⅰ. 大… Ⅱ. 托… Ⅲ. 数据处理－英文 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2017）第 233095 号

本书版权登记号：图字 01-2016-4290

Authorized reprint from the English language edition, entitled *Big Data Fundamentals: Concepts, Drivers, and Techniques*, 1E, Thomas Erl, Wajid Khattak, Paul Buhler, published by Pearson Education, Inc., Copyright © 2016.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc.

English language edition published by China Machine Press/Huazhang Co., Copyright © 2017.

本书英文影印版由 Pearson Education Inc. 授权机械工业出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

此影印版仅限于中华人民共和国境内（不包括香港、澳门特别行政区及台湾地区）销售发行。

本书封面贴有 Pearson Education（培生教育出版集团）激光防伪标签，无标签者不得销售。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：曲 熠 责任校对：李秋荣
印 刷：北京文昌阁彩色印刷有限责任公司 版 次：2017 年 10 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：14
书 号：ISBN 978-7-111-58098-0 定 价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

版权所有·侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所 韩光/邹晓东

出版者的话

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的优势，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brian W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力相助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专门为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com
电子邮件：hzjsj@hzbook.com
联系电话：（010）88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037



华章教育

华章科技图书出版中心

致　谢

按姓氏字母排序：

Allen Afuah，密歇根大学罗斯商学院

Thomas Davenport，巴布森学院

Hugh Dubberly，Dubberly 设计工作室

Joe Gollner，Gnostyx 研究公司

Dominic Greenwood，Whitestein 技术公司

Gareth Morgan，约克大学加里斯摩根商学院

Peter Morville，Semantic 工作室

Michael Porter，哈佛商学院战略与竞争研究所

Mark von Rosing，LEADing Practice

Jeanne Ross，麻省理工学院斯隆管理学院信息系统研究中心

Jim Sinur，Flueresque 公司

John Sterman，麻省理工学院斯隆管理学院 MIT 系统动力学组

本书以大数据科学专家认证（BDSCP）课程为基础，特别感谢 Arcitura 教育公司以及相关的大数据研究与开发团队的辛苦工作。

作者简介

Thomas Erl

Thomas Erl 是 IT 畅销书作者，Arcitura 教育公司的创始人，Prentice Hall 出版社“Thomas Erl 的服务技术丛书”的编辑。他的书发行量超过 200 000 册，成为国际畅销书，并且已经获得多个重要 IT 组织成员的正式认可，例如，IBM、Microsoft、Oracle、Intel、Accenture、IEEE、HL7、MITRE、SAP、CISCO、HP 等。作为 Arcitura 公司的 CEO，Thomas 领导研发了国际公认的大数据科学专家认证（BDSCP）、云专家认证（CCP）与 SOA 专家认证（SOACP）的课程大纲，设立了一系列正式的、与厂商无关的工业认证，全球已有数千 IT 从业人员获得了这些认证。Thomas 还作为演讲家与教育家，在 20 多个国家进行过巡回演讲。Thomas 已经在诸多出刊物上发表过 100 多篇文章和访谈，包括《华尔街日报》与《CIO 杂志》。

Wajid Khattak

Wajid Khattak 是 Arcitura 教育公司的大数据研究者与教育者。他的研究领域包括大数据工程与架构、数据科学、机器学习、分析学与 SOA。此外，他在商务智能报告解决方案与 GIS 方面有着丰富的 .NET 软件开发经验。

Wajid 于 2003 年在英国伯明翰城市大学获得软件工程学士学位，于 2008 年在该校以杰出的成绩获得软件工程与安全硕士学位。另外，Wajid 还获得了 MCAD & MCTS（Microsoft）、SOA 架构师、大数据科学家、大数据工程师以及大数据研究顾问（Arcitura）认证。

Paul Buhler

Paul Buhler 博士是一位经验丰富的 IT 专家，他在商业公司、政府机构和学校均有过从业经验。在面向服务的计算概念、技术和实现方法领域，他是一位受人尊敬的研究者、实践者与教育者。他在 XaaS 领域的研究已经延伸到了云、大数据与万物互联网（IoE）。目前他的研究兴趣是通过权衡响应式设计原则与基于目标的执行方式，减少业务策略与流程执行之间的差距。

作为 Modus21 的首席科学家，Paul Buhler 博士根据当前业务架构与流程执行框架的发展趋势调整企业的战略布局。目前，他还是查尔斯顿学院的合作教授，负责本科生与硕士生计算机科学课程的教学工作。Paul Buhler 博士在南卡罗来纳大学获得计算机工程博士学位，在约翰霍普金斯大学获得计算机科学硕士学位，在塞特多大学获得计算机科学学士学位。

Contents

PART I: THE FUNDAMENTALS OF BIG DATA

CHAPTER 1: Understanding Big Data 3

Concepts and Terminology 5
Datasets 5
Data Analysis 6
Data Analytics 6
Descriptive Analytics *8*
Diagnostic Analytics *9*
Predictive Analytics *10*
Prescriptive Analytics *11*
Business Intelligence (BI) 12
Key Performance Indicators (KPI) 12
Big Data Characteristics 13
Volume 14
Velocity 14
Variety 15
Veracity 16
Value 16
Different Types of Data 17
Structured Data 18
Unstructured Data 19
Semi-structured Data 19
Metadata 20
Case Study Background 20
History 20
Technical Infrastructure and Automation Environment 21
Business Goals and Obstacles 22
Case Study Example 24
Identifying Data Characteristics 26
Volume *26*
Velocity *26*

Variety 26
Veracity 26
Value 27
Identifying Types of Data 27

CHAPTER 2: Business Motivations and Drivers for Big Data Adoption 29

Marketplace Dynamics 30
Business Architecture 33
Business Process Management 36
Information and Communications Technology 37
Data Analytics and Data Science 37
Digitization 38
Affordable Technology and Commodity Hardware 38
Social Media 39
Hyper-Connected Communities and Devices 40
Cloud Computing 40
Internet of Everything (IoE) 42
Case Study Example 43

CHAPTER 3: Big Data Adoption and Planning Considerations 47

Organization Prerequisites 49
Data Procurement 49
Privacy 49
Security 50
Provenance 51
Limited Realtime Support 52
Distinct Performance Challenges 53
Distinct Governance Requirements 53
Distinct Methodology 53
Clouds 54

Big Data Analytics Lifecycle 55
Business Case Evaluation 56
Data Identification 57
Data Acquisition and Filtering 58
Data Extraction 60
Data Validation and Cleansing 62
Data Aggregation and Representation 64
Data Analysis 66
Data Visualization 68
Utilization of Analysis Results 69
Case Study Example 71
Big Data Analytics Lifecycle 73
Business Case Evaluation 73
Data Identification 74
Data Acquisition and Filtering 74
Data Extraction 74
Data Validation and Cleansing 75
Data Aggregation and Representation 75
Data Analysis 75
Data Visualization 76
Utilization of Analysis Results 76

CHAPTER 4: Enterprise Technologies and Big Data Business Intelligence 77

Online Transaction Processing (OLTP) 78
Online Analytical Processing (OLAP) 79
Extract Transform Load (ETL) 79
Data Warehouses 80
Data Marts 81
Traditional BI 82
Ad-hoc Reports 82
Dashboards 82
Big Data BI 84
Traditional Data Visualization 84
Data Visualization for Big Data 85

Case Study Example 86
Enterprise Technology 86
Big Data Business Intelligence 87

PART II: STORING AND ANALYZING BIG DATA

CHAPTER 5: Big Data Storage Concepts 91

Clusters 93
File Systems and Distributed File Systems 93
NoSQL 94
Sharding 95
Replication 97
Master-Slave 98
Peer-to-Peer 100
Sharding and Replication 103
Combining Sharding and Master-Slave Replication 104
Combining Sharding and Peer-to-Peer Replication 105
CAP Theorem 106
ACID 108
BASE 113
Case Study Example 117

CHAPTER 6: Big Data Processing Concepts 119

Parallel Data Processing 120
Distributed Data Processing 121
Hadoop 122
Processing Workloads 122
Batch 123
Transactional 123
Cluster 124

Processing in Batch Mode 125
Batch Processing with MapReduce 125
Map and Reduce Tasks 126
Map *127*
Combine *127*
Partition *129*
Shuffle and Sort *130*
Reduce *131*
A Simple MapReduce Example 133
Understanding MapReduce Algorithms 134
Processing in Realtime Mode 137
Speed Consistency Volume (SCV) 137
Event Stream Processing 140
Complex Event Processing 141
Realtime Big Data Processing and SCV 141
Realtime Big Data Processing and MapReduce 142
Case Study Example 143
Processing Workloads 143
Processing in Batch Mode 143
Processing in Realtime 144

CHAPTER 7: Big Data Storage Technology 145

On-Disk Storage Devices 147
Distributed File Systems 147
RDBMS Databases 149
NoSQL Databases 152
Characteristics *152*
Rationale *153*
Types *154*
Key-Value *156*
Document *157*
Column-Family *159*
Graph *160*
NewSQL Databases 163
In-Memory Storage Devices 163
In-Memory Data Grids 166
Read-through *170*

Write-through 170
Write-behind 172
Refresh-ahead 172
In-Memory Databases 175
Case Study Example 179

CHAPTER 8: Big Data Analysis Techniques 181

Quantitative Analysis 183
Qualitative Analysis 184
Data Mining 184
Statistical Analysis 184
A/B Testing 185
Correlation 186
Regression 188
Machine Learning 190
Classification (Supervised Machine Learning) 190
Clustering (Unsupervised Machine Learning) 191
Outlier Detection 192
Filtering 193
Semantic Analysis 195
Natural Language Processing 195
Text Analytics 196
Sentiment Analysis 197
Visual Analysis 198
Heat Maps 198
Time Series Plots 200
Network Graphs 201
Spatial Data Mapping 202
Case Study Example 204
Correlation 204
Regression 204
Time Series Plot 205
Clustering 205
Classification 205

APPENDIX A: Case Study Conclusion 207

目　录

第一部分　大数据基础

第 1 章　理解大数据 …… 3

概念与术语 …… 5
数据集 …… 5
数据分析 …… 6
数据分析学 …… 6
描述性分析 …… 8
诊断性分析 …… 9
预测性分析 …… 10
规范性分析 …… 11
商务智能 …… 12
关键绩效指标 …… 12
大数据特征 …… 13
容量 …… 14
速率 …… 14
多样性 …… 15
真实性 …… 16
价值 …… 16
不同数据类型 …… 17
结构化数据 …… 18
非结构化数据 …… 19
半结构化数据 …… 19
元数据 …… 20
案例学习背景 …… 20
历史背景 …… 20
技术基础和自动化环境 …… 21
商业目标和障碍 …… 22
案例学习 …… 24
确定数据特征 …… 26
容量 …… 26
速率 …… 26
多样性 …… 26
真实性 …… 26
价值 …… 27
确定数据类型 …… 27

第 2 章　采用大数据的商业动机与驱动 …… 29

市场动态 …… 30
业务架构 …… 33
业务流程管理 …… 36
信息与通信技术 …… 37
数据分析与数据科学 …… 37
数字化 …… 38
开源技术与商用硬件 …… 38
社交媒体 …… 39
超连通社区与设备 …… 40
云计算 …… 40
万物互联网 …… 42
案例学习 …… 43

第 3 章　大数据采用及规划考虑 …… 47

组织的先决条件 …… 49
数据获取 …… 49
隐私性 …… 49
安全性 …… 50
数据来源 …… 51
有限的实时支持 …… 52
不同的性能挑战 …… 53
不同的管理需求 …… 53
不同的方法论 …… 53
云 …… 54
大数据分析的生命周期 …… 55

商业案例评估 ……56
数据标识 ……57
数据获取与过滤 ……58
数据提取 ……60
数据验证与清理 ……62
数据聚合与表示 ……64
数据分析 ……66
数据可视化 ……68
分析结果的使用 ……69
案例学习 ……71
大数据分析的生命周期 ……73
商业案例评估 ……73
数据标识 ……74
数据获取与过滤 ……74
数据提取 ……74
数据验证与清理 ……75
数据聚合与表示 ……75
数据分析 ……75
数据可视化 ……76
分析结果的使用 ……76

第 4 章　企业级技术与大数据商务智能 ……77

联机事务处理 ……78
联机分析处理 ……79
抽取、转换和加载技术 ……79
数据仓库 ……80
数据集市 ……81
传统商务智能 ……82
即席报表 ……82
仪表板 ……82
大数据商务智能 ……84
传统数据可视化 ……84
大数据的数据可视化 ……85
案例学习 ……86
企业技术 ……86
大数据商务智能 ……87

第二部分　存储和分析大数据

第 5 章　大数据存储的概念 ……91

集群 ……93
文件系统和分布式文件系统 ……93
NoSQL ……94
分片 ……95
复制 ……97
主从式复制 ……98
对等式复制 ……100
分片和复制 ……103
结合分片和主从式复制 ……104
结合分片和对等式复制 ……105
CAP 定理 ……106
ACID ……108
BASE ……113
案例学习 ……117

第 6 章　大数据处理的概念 ……119

并行数据处理 ……120
分布式数据处理 ……121
Hadoop ……122
处理工作量 ……122
批处理型 ……123
事务型 ……123
集群 ……124
批处理模式 ……125
MapReduce 批处理 ……125
Map 和 Reduce 任务 ……126
映射 ……127
合并 ……127
分区 ……129
洗牌和排序 ……130
归约 ……131

MapReduce 的简单实例 …… 133
理解 MapReduce 算法 …… 134
实时模式处理 …… 137
SCV 原则 …… 137
事件流处理 …… 140
复杂事件处理 …… 141
大数据实时处理与 SCV …… 141
大数据实时处理与 MapReduce …… 142
案例学习 …… 143
处理工作量 …… 143
批处理模式处理 …… 143
实时模式处理 …… 144

第 7 章　大数据存储技术 …… 145

磁盘存储设备 …… 147
分布式文件系统 …… 147
RDBMS 数据库 …… 149
NoSQL 数据库 …… 152
特征 …… 152
理论基础 …… 153
类型 …… 154
键 – 值存储 …… 156
文档存储 …… 157
列簇存储 …… 159
图存储 …… 160
NewSQL 数据库 …… 163
内存存储设备 …… 163
内存数据网格 …… 166
同步读 …… 170
同步写 …… 170
异步写 …… 172
异步刷新 …… 172
内存数据库 …… 175
案例学习 …… 179

第 8 章　大数据分析技术 …… 181

定量分析 …… 183
定性分析 …… 184
数据挖掘 …… 184
统计分析 …… 184
A/B 测试 …… 185
相关性分析 …… 186
回归性分析 …… 188
机器学习 …… 190
分类（有监督的机器学习） …… 190
聚类（无监督的机器学习） …… 191
异常检测 …… 192
过滤 …… 193
语义分析 …… 195
自然语言处理 …… 195
文本分析 …… 196
情感分析 …… 197
视觉分析 …… 198
热点图 …… 198
时间序列图 …… 200
网络图 …… 201
空间数据制图 …… 202
案例学习 …… 204
相关性分析 …… 204
回归性分析 …… 204
时间序列图 …… 205
聚类 …… 205
分类 …… 205

附录 A　案例结论 …… 207



Part I

The Fundamentals of Big Data

Chapter 1 Understanding Big Data

Chapter 2 Business Motivations and Drivers for Big Data Adoption

Chapter 3 Big Data Adoption and Planning Considerations

Chapter 4 Enterprise Technologies and Big Data Business Intelligence

Big Data has the ability to change the nature of a business. In fact, there are many firms whose sole existence is based upon their capability to generate insights that only Big Data can deliver. This first set of chapters covers the essentials of Big Data, primarily from a business perspective. Businesses need to understand that Big Data is not just about technology—it is also about how these technologies can propel an organization forward.

Part I has the following structure:

- Chapter 1 delivers insight into key concepts and terminology that define the very essence of Big Data and the promise it holds to deliver sophisticated business insights. The various characteristics that distinguish Big Data datasets are explained, as are definitions of the different types of data that can be subject to its analysis techniques.
- Chapter 2 seeks to answer the question of why businesses should be motivated to adopt Big Data as a consequence of underlying shifts in the marketplace and business world. Big Data is not a technology related to business transformation; instead, it enables innovation within an enterprise on the condition that the enterprise acts upon its insights.
- Chapter 3 shows that Big Data is not simply "business as usual," and that the decision to adopt Big Data must take into account many business and technology considerations. This underscores the fact that Big Data opens an enterprise to external data influences that must be governed and managed. Likewise, the Big Data analytics lifecycle imposes distinct processing requirements.
- Chapter 4 examines current approaches to enterprise data warehousing and business intelligence. It then expands this notion to show that Big Data storage and analysis resources can be used in conjunction with corporate performance monitoring tools to broaden the analytic capabilities of the enterprise and deepen the insights delivered by Business Intelligence.

Big Data used correctly is part of a strategic initiative built upon the premise that the internal data within a business does not hold all the answers. In other words, Big Data is not simply about data management problems that can be solved with technology. It is about business problems whose solutions are enabled by technology that can support the analysis of Big Data datasets. For this reason, the business-focused discussion in Part I sets the stage for the technology-focused topics covered in Part II.